哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

Chips Computers Crops

2014 International Workshop and Summer School on "Biological Big Bytes"
a Satellite Meeting of the International Workshop/Summer school on Crops, Chips & Computers

Harbin Institute of Technology, Harbin, China

August 17th to 23th, 2014

WELCOME TO C3-B3'2014

Dear Colleagues and Friends:

Welcome to C3-B3'2014 at the Harbin Institute of Technology, Harbin, China, August 17th-23th,2014!

This workshop and summer school on Biological Big Bytes (B3) is a satellite meeting of our traditional annual international workshop and summer school on "Crops, Chips and Computers" (C3), aiming to teach best practice in the application of state-of-the-art information technology and computational techniques to big data research in life sciences. The workshop and summer school will cover interdisciplinary science, in a rapidly evolving field, bringing together ideas from the fields of mathematics, bioinformatics, omics and computer sciences. Because of these attractions, we expect the school to be of interest to a range of students from all academic levels, from post-graduate students to postdoctoral workers to junior academics.

Organizers

Dr. Andrew Harrison (University of Essex)

Dr. Christian Klukas (Leibniz Institute of Plant Genetics and Crop Plant Research)

Dr. Dechang Xu (Harbin Institute of Technology)

Dr. Huaqin He (Fujian Agriculture and Forestry University University)

Dr. Hugh Shanahan (Royal Holloway, Unversity of London)

Dr. Ling-ling Chen (Huazhong Agricultural University)

Dr. Ming Chen (Zhejiang University)

Dr. Qingfeng Chen (Guangxi University)

Dr. Sean May (The University of Nottingham)

Dr. Yongsheng Chen (Inner Mongolian University for the Nationalities)

Dr. Ziding Zhang (China Agricultural University)

Funding and Partners

The 2014 International Workshop and Summer School on Biological Big Bytes is sponsored by Harbin Institute of Technology. It is jointly organized by the Department of Bioinformatics, College of Life Sciences, Zhejiang University, and the Departments of Mathematical Sciences & Biological Sciences, University of Essex. It is hosted by Harbin Institute of Technology, Harbin, China.

Local organizer

Dr. Nanqi Ren, (Harbin Institute of Technology), Chair

Dr. Dechang Yu (Harbin Institute of Technology), Co-Chair

Dr. Ming Du (Harbin Institute of Technology)

Dr. Fei Liu (Harbin Institute of Technology)

Contact

Email: All questions related to the C3'2014 can be sent to

Dr. Dechang Xu (dcxu@hit.edu.cn) (Harbin Institute of Technology) or

Dr. Ming Chen (mchen@zju.edu.cn) (Zhejiang University) or

Dr. Andrew Harrison (harry@essex.ac.uk) (University of Essex)

Tel: +(0)451-86282904

Address: School of Food Science & Engineering, Harbin Institute of Technology, Harbin 150090, P. R. China

Travel

By Air-plane

From Airport to Harbin Institute of Technology:

You may come to Harbin via Beijing / Shanghai.

Taxi: Taxi from airport to the 2nd campus costs about 120 RMB

Airport Bus No.2: costs 20 RMB

By Train:

There are three railway stations in Harbin.

From Railway Station:

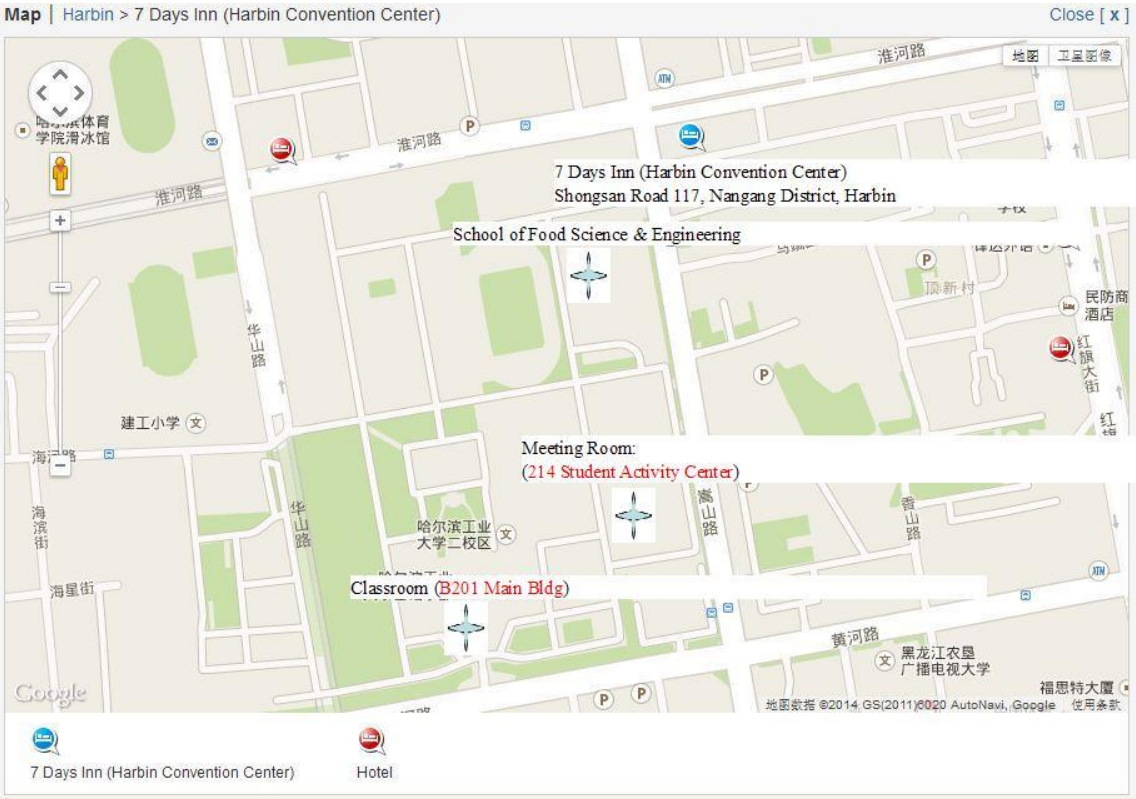
Taxi: Taxi from Harbin Station to the 2nd campus costs about 20 RMB.

Bus: by Bus No. 89 costs 1 RMB

Hotel

7 Days Inn (Harbin Convention Center)

Shongsan Road 117, Nangang District, Harbin



Programme

Sunday 17th August 2014

Registration Harbin Institute of Technology

17:30~ ***Dinner***

Monday 18th August 2014

Registration Harbin Institute of Technology Classroom (B201 Main Bldg)

08:30~09:00 ***Introduction*** Ming Chen & Andrew Harrison & Dechang Xu

An overview of what needs to be done at the school and the expected work load on each day.

Team grouping

Section 1 (Lectures)

09:00~10:15 ***Lecture 1*** history of science, technology and data

Lecturer Andrew Harrison

10:30~11:45 ***Lecture 2*** practicalities of scientific data and the wider zeitgeist of data.

Lecturer Andrew Harrison

12:00~13:30 ***Lunch***

Section 2 (Practical)

14:30~17:30 ***Practice 1*** practicalities of scientific data and the wider zeitgeist of data.

Lecturer Andrew Harrison

18:00~ ***Dinner***

Tuesday 19th August 2014

Section 3 (Lectures)

09:00~10:00 ***Lecture 3*** Files vs Databases / SQL vs NoSQL.

Lecturer Adam

10:10~11:00 ***Lecture 4*** An introduction to SQL & Database Design. An introduction to sed, awk and grep

Lecturer Andrew Harrison

12:00~13:30 ***Lunch***

Section 4 (Practical)

14:30~17:30 ***Practice 2*** File naming, versioning and backing up. Data and metadata preparation and publishing, persistent identifiers and citation, linking to articles etc.

Practical Louise

18:00~ ***Dinner***

Wednesday 20th August 2014

Workshop Day

Meeting Room 214 Student Activity Center

Section 9 (Chair: Huaqin He)

08:30~09:10 ***Opening Ceremony***

Speaker President

Group photo

09:10~9:40 ***Speaker 1*** Hongzhi Wang

Title Algorithm Design and Analysis for Big Data

09:40~10:10 ***Speaker 2*** Adam Carter
Title The EUDAT Collaborative Data Infrastructure
10:10~10:30 ***Coffee/Tea break***

Section 10 (Chair: Ming Chen)
10:30~10:55 ***Speaker 3*** Huaqin He
Title Genome-wide identification and function of heat-shock proteins in rice
10:55~11:20 ***Speaker 4*** Xiaobao Dong / Ziding Zhang
Title Revealing Shared and Distinct Network Organizations of Arabidopsis Immune Responses by Integrative Analysis
11:20~11:45 ***Speaker 5*** Andrew Harrison
Title Mind-altering microbes
11:45~12:10 ***Speaker 6*** Jingjing Wang / Ming Chen
Title The roles of cross-talking epigenetic patterns in Arabidopsis thaliana
12:30~13:30 ***Lunch***

Section 11 (Chair: Ziding Zhang)
14:30~14:55 ***Speaker 7*** Professionalising digital data curation skills in Universities
Title Louise Corti
14:55~15:20 ***Speaker 8*** Fei Liu
Title Colored Petri Nets for Multiscale Systems Biology
15:20~15:45 ***Speaker 9*** Lifeng Xu
Title Simulating genotype-phenotype interactions using Functional-Structural plant models
15:45~16:10 ***Speaker 10*** Bjorn Sommer
Title CELLmicrocosmos Workshop: Cell Modeling at Themesoscopic and Functional Level
16:10~16:30 ***Coffee/Tea break***

Section 12: iGEM MeetUp (Chair: Dechang Xu)
16:30~16:55 ***Speaker 11*** Kang Yang
Title Aspergillus niger is coming!
16:55~17:20 ***Speaker 12*** Zhiqi Liu
Title E.conan detective
17:20~17:45 ***Speaker 13*** Yifan Wu
Title Dioxin Detective Y
18:00~20:00 ***Dinner & Discussion***

Thursday 21th August 2014

Section 5 (Lectures)
09:00~10:30 ***Lecture 5*** Big Science with Big Data needs Data Infrastrucutres - Data has value - Data reuse - Data sharing - Distributed data Persistent Identifiers and Metadata (bringing out synergies with Louise's material).
Lecturer Adam
10:50~11:30 ***Lecture 6***
Lecturer
12:00~13:30 ***Lunch***

Section 6 (Practical)
14:30~16:00 ***Practice 3*** Data publishing practical.
Practical Louise

16:00~17:00 ***Practice 4***
. ***Lecturer***
18:00~ ***Dinner***

Friday 22th August 2014

	Section 7 (Lectures)
09:00~10:30	***Lecture 7*** Models for Processing Big Data - Moving Compute to the Data - Workflows - Data Intensive Compute Architectures Models for Processing Big Data - MapReduce (includes practical session)
	Lecturer Adam
10:50~11:30	***Lecture 8*** Practice and Discussion.
	Lecturer
12:00~13:30	***Lunch***
	Section 8 (Practical)
14:30~17:30	***Practice 5*** Discussion about careers in data Student presentations.
	Practical
18:00~	***Dinner***

Saturday 23th August 2014

Tour

Sunday 24 th August 2014

9:30~11:40	***Close session & Awards ceremony***
	Science in the UK, Germany and China, and moving among countries; career structures in these countries. We expect to recruit European students who may be considering spending part of their careers in China and we hope some of the Chinese students similarly look to spend part of their careers in Europea.
12:00~	***Lunch & Farewell***

Workshop abstracts

- **Title：** Algorithm Design and Analysis for Big Data

Hongzhi Wang Assoc. Professor, *School of Computer Science & Technology, HIT*

Abstract: Big data bring challenges for algorithm design and analysis. Algorithms for big data analysis are required to handle the volume and velocity features. Meeting these difficulties, some new algorithms design and analysis techniques are proposed. This talk shows the widely applications of big data algorithms and gives an overview of techniques for big data algorithm design and analysis including sub-linear algorithms, I/O efficient algorithms, parallel algorithms as well as crowd sourcing algorithms.

- **Title：** The EUDAT Collaborative Data Infrastructure

Adam Carter, *EPCC, The University of Edinburgh*

Abstract： EUDAT's goal is to build a Collaborative Data Infrastructure (CDI) as a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. The CDI will allow researchers to share data within and between communities and enable them to carry out their research effectively. Our mission is to provide a solution that will be affordable, trustworthy, robust, persistent, open and easy to use.

In this talk I will discuss the progress that EUDAT has made towards these goals, and I'll describe the EUDAT services that are already up and running.

- **Title**: Genome-wide identification and function of heat-shock proteins in rice

Yongfei Wang1, Xinhai Chen1, Shoukai Lin1,2*, Qi Song1, Huan Tao1, Jian Huang1, Yuqin Ke1, Shufu Que1, **Huaqin He**1*

1 College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China.

2 Putian University, Putian, Fujian, China

Abstract： Heat shock proteins (Hsps) have been known as a class of proteins that perform a fundamental role in protecting plants against abiotic stresses, such as drought, salinity, extremely low and high temperatures. However, rare Hsps have been reported and little consideration has been given to the role of Hsps in rice in response to abiotic stress. Therefore, we conducted this research to identify heat shock proteins and reveal the role of sHsps in rice in response to abiotic stresses.

First, we combined an orthology-based approach with expression association data to screen rice Hsps, the expression patterns of which strongly correlate with those of heat responsive probe-sets. Twenty-seven Hsp candidates were verified, including 12 small Hsps, six Hsp70s, three Hsp60s, three Hsp90s, and three clpB/Hsp100s. Then, using a combination of interolog and expression profile-based methods, we identified 430

interactorsofHsp70 family members in rice, and validated the interactions by co-localization and function-based methods. Subsequently, 13 interacting domains and 28 target motifs were over-represented in Hsp70sinteractors. Twenty-four GO terms of biological processes and five GO terms of molecular functions were enriched in the positive Hsp70 interactors, which expression levels were positively associated with Hsp70s. Analyzing the interaction network of Hsp70s, we found that Hsp70s were involved in macromolecular translocation, carbohydrate metabolism, innate immunity, photosystem II repair and regulation of kinase activities.

Hsp26.7, Hsp23.2, Hsp17.9A, Hsp17.4 and Hsp16.9A were up-regulated in rice (Orya sativa L. & Nipponbare) during seedling and anthesis stages in response to heat stress. Subsequently, the expression level of five sHsps in the heat-tolerant rice cultivar, Co39, were all significantly higher than that in the heat-susceptible rice cultivar, Azucena, indicating that the expression level of these five sHsps was positive related to the ability of rice plants to avoid heat stress. Thus, the expression level of these five sHsps can be regarded as bio-markers for screening rice cultivars with different abilities to avoid heat stress. Hsp18.1, Hsp17.9A, Hsp17.7 and Hsp16.9A, in the three rice cultivars under heat stress were found to be involved in one protein complex by Native-PAGE, and the interactions of Hsp18.1 and Hsp 17.7, Hsp18.1 and Hsp 16.9A, Hsp17.7 and Hsp16.9A were further validated by yeast 2-hybridization. Pull down assay also confirmed the interaction between Hsp17.7 and Hsp16.9A in rice under heat stress. In conclusion, the up-regulation of the 5 sHsps is a key step for rice to tolerate heat stress, after that some sHsps are assembled into a large hetero-oligomeric complex.

Moreover, we constructed a Protein-Protein Interaction Predictor (PPIP) and a Rice Gene Expression Profile database for experimental biologists. These tools could be available at http://bioinformatics.fafu.edu.cn/yfwang/PPIP/ppip.php and http://bioinformatics.fafu.edu.cn/ yfwang/array/index.php respectively.

Keywords: Rice (Oryza sativa L.), Heat shock proteins, Expression, Interactors

ACKNOWLEDGMENT
This work was supported by the Natural Science Foundation of China and Fujian (grant nos. 31270454, 61163047 and 2013J01077), a grant from the Education Department of Fujian (grant no. JA10103) and the Key Program of Ecology in Fujian (grant nos. 0608507 and 6112C0600).

- **Title**: Revealing Shared and Distinct Network Organizations of Arabidopsis Immune Responses by Integrative Analysis

Xiaobao Dong* and Ziding Zhang

State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

* Speaker (Email: x.b.dong@cau.edu.cn)

Abstract: Pattern-triggered immunity (PTI) and effector-triggered immunity (ETI) are two main forms of plant immune response to counter pathogen invasion. Up to now, genome-wide regulatory network organization principles leading to quantitative differences between PTI and ETI remain elusive. We combined an advanced

machine learning method and modular network analysis to systematically characterize the organization principles of Arabidopsis PTI and ETI. We report our major findings from three network resolutions. At a single network node/edge level, we ranked important genes and gene interactions for immune response and successfully identified many known immune regulators for PTI and ETI, respectively. Topological analysis showed that important gene interactions tend to link network modules. At a subnetwork level, we identified a subnetwork shared by PTI and ETI, which covers 1159 genes and 1289 interactions. In addition to being enriched with interactions linking network modules, it is also a hotspot attacked by pathogen effectors. The subnetwork likely represents a core component to coordinate multiple biological processes in the transition from development to defense. Finally, we constructed modular network models for PTI and ETI to explain the quantitative differences from the global network architecture. Our results show defense modules appeared to be interdependently connected in PTI, but independently connected in ETI, providing an explanation for the robustness of ETI to genetic mutations and effector attacks. Taken together, the multiscale comparisons between PTI and ETI provide a systems biology point of view to understand plant immunity, and highlight the coordination among network modules to establish a robust immune response.

Keywords: pattern-triggered immunity, effector-triggered immunity, regulatory network, subnetwork, network module

- **Title**: Mind-altering microbes

Andrew Harrison *University of Essex*

Abstract: We are moving through a paradigm shift in how we view the relationship between Homo Sapiens and microbes - we may be best considered as a mammalian-microbe symbiotic ecosystem. This changing view will impact upon many fields, including psychology and mental health. It is increasingly clear that the Western lifestyle is leading to a dramatic change in our microbiota. This may have causal implications for the dramatic rises we are witnessing in obesity, autoimmune diseases and mental disorders such as autism. It will further impact upon how we care for an ageing population, overcome malnutrition, and deal with infectious diseases in the post-antibiotic era.

- **Title**: The roles of cross-talking epigenetic patterns in *Arabidopsis thaliana*

Jingjing Wang[1, 2,*], Xianwen Meng[1], Chunhui Yuan[1], Andrew P. Harrison[3] and Ming Chen[1, 2]

[1]*Department of Bioinformatics, the State Key Laboratory of Plant Physiology and Biochemistry, Institute of Plant Science; College of Life Sciences, Zhejiang University, Hangzhou 310058, P.R. China,* [2]*James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310058, P. R. China,* [3]*Department of Mathematical Sciences and School of Biological Sciences, University of Essex, Colchester, Essex CO4 3SQ, UK*

* Speaker (Email: jingjingw@zju.edu.cn)

Abstract: Epigenetic marks, including histone modifications, DNA methylation and histone variants, play a key role in determining transcriptional outcomes. The availability of the high-throughput sequencing data allows us to characterize the genome-wide distribution patterns of epigenetic marks, which have improved our understanding of the relationship between chromatin modification/structure and genes expression. More and more studies have found that epigenetic mechanisms have been implicated in many biological processes, such as Transposon Element silencing and long-range interactions within the genome. Moreover, those studies have found there are different layers of cross-talk and interplay among epigenetic marks to be involved in chromatin-dependent processes. Here, we take advantage of evidence from recent genetic, biochemical and high-throughput sequencing studies to explore the complex cross-talk between different epigenetic patterns, and address how these epigenetic marks interplay with each other and then affect gene transcription. Finally, we build a cross-talk network and analyze the characteristic of the network. Comprehensive cross-talk network of epigenetic marks will help in fully understanding the functional roles and biological impact of epigenetic regulation.

Keywords: histone modification; DNA methylation; histone variant; cross-talk

- **Title：** 'Professionalising digital data curation skills in Universities'

Louise Corti *UK Data Archive*

Abstract： Researchers' responsibilities towards their research data are set to change across all domains of scientific endeavour. Research funders are increasingly mandating open access to research data; governments internationally are demanding transparency in research; the economic climate is requiring much greater re-use of data; and fear of data loss calls for more robust information security practices. All these factors mean that researchers need to improve, enhance and professionalize their research data management skills to meet the challenge of producing the highest quality shareable and reusable research outputs in a responsible and efficient way. The promotion of these skills offers a strategic contribution to the world's research capacity building programmes. How can we go about professionalising our data management skills? I take a guided tour of various efforts in UK universities to help make this agenda happen.

- **Title：** Colored Petri Nets for Multiscale Systems Biology

Fei Liu *Control and Simulation Center, HIT*
Email: liufei@hit.edu.cn

Abstract: Systems biology aims to understand the behavior of a biological system at the system level by means of investigating the behavior and interactions of all the components in the system. Due to the ability to produce data of one and the same phenomenon at different scales, the modeling of biological systems is moving from single scales to multiple scales, e.g., from the molecular scale to the cell, tissue, and even the whole organism

level. In this report, we will present a colored Petri net pproach to modeling and analyzing multiscale systems biology.Specifically, in this report you will see what are colored Petri nets, what challenges of multiscale systems biology can be tackled by colored Petri ets, a colored Petri net framework for multiscale systems biology, and a couple of case studies.

- **Title：** Simulating genotype-phenotype interactions using Functional-Structural plant models

Lifeng Xu *College of Computer Science & Technology, Zhejiang University of Technology, 310023 Hangzhou, China*
Email: lfxu@zjut.edu.cn

Crop breeders and breeding scientists presently face a multitude of problems and challenges: Cereals, namely rice, maize and wheat, constitute the basic staple foods and the demand for them is rising with an ever-increasing human population. Therefore, understanding plant growth and morphology is particularly of great importance to agronomy. Modelling efforts within the Functional-Structural Plant Model (FSPM) domain are mostly concerned with the acquisition, transport and use of matter and energy from sources to sinks through pathways dictated by plant structure, such as light, carbon, water and soil minerals, and how this affects growth and morphology of the resulting plants. Genotype-phenotype models of crops in the form of spatially explicit models of morphogenetic development, based on the interaction of compartmentalized physiological processes, and coupled with quantitative genetic information, can be used to better understand crop systems and ultimately to aid breeding of new crop plant varieties. In order to illustrate our approach of simulating genotype-phenotype interactions, an FSPM of rice will be described here. The rice model can simulate growth and morphology of rice plant from germination to seed maturity, in combination with selected ecophysiological processes including photosynthesis and sink functions based on a common assimilate pool. Furthermore, it was linked to a quantitative genetic model, thereby simulating the genetic reproduction processes to form new rice population. The rice FSPM reproduces plant architecture, morphology and the QTLs for certain traits (in this case: plant height) of rice from germination to maturity. The potential use of the model with further extension would be as a tool in domains of FSPM education and scientific research.

- **Title：** SUBCELLULAR VISUALIZATION AND LOCALIZATION OF PROTEIN-RELATED DATA

Björn Sommer *Bio-/Medical Informatics Department, University Bielefeld, Bielefeld, Germany*
E-mail: bjoem@CELLmicrocosmos.org

Molecular databases such as Reactome and KEGG support the visualization and analysis of biomedical networks [COWH11, KGSF12]. These pathways usually consist of enzymes, substrates and products, and their corresponding reaction processes.Websites of both databases provide 2D network visualizations which are often enriched with the information where specific genes or proteins are located. These images are curated and the corresponding subcellular localizations are usually reliable.

But problems arise if biomedical networks are (semi-)automatically generated based on heterogeneous database information. For example, if a set of proteins – e.g. correlated with a specific disease – is found by combining information from different databases, consistent information regarding the subcellular localization is usually missing. To obtain the localization information, again various databases can be consulted. But the problem is now to obtain reasonable localizations in the context of the newly created biological network. For this purpose, the CELLmicrocosmos 4 PathwayIntegration was developed:

http://Cm4.CELLmicrocosmos.org

Using subcellular localization charts it is possible to compare, analyze and predict the protein localization based on a number of established databases [SKDA13]. Direct links to the original sources enable the validation of the localization predictions. Moreover, the results can be mapped into a virtual cell 3D environment which is directly connected to the 2D visualization of the biomedical pathway. Based on the jSBML library, protein-related networks can be imported from various sources[DRDD11].

iGEM MeetUp

- **Title：** Aspergillus niger is coming !

Kang Yang *Northeast Agricultural University*

Abstract: Aspergillus niger, a species of filamentous fungi, is universally acknowledged of having long been investigated for industry production of metabolites and enzymes. Meanwhile, with the growth of the biotechnology industry, it has been gradually employed in areas as diverse as medicine, agriculture, and basic science that influence our everyday lives. Taking advance of the expression system in Aspergillus niger, our team successfully accomplished “Visual operation system of continuous gene replacement” and “From Red to Cyan ”, respectively means targeted gene’s location at specific sites and transformation accelerator from lycopene to astaxanthin. The meaning of our work, on one hand, relied on that we established a new visual operation system assisting us to make sure where is the DNA. On the other hand, we designed one/two fusion protein contributing to the transformation accelerator of metabolism.

- **Title：** E.conan detective

Zhiqi Liu ***Northeast Forestry University***

Abstract： There is no doubt that heavy metal contamination has detrimental effects on our environment and our health. For the sake of solving this challenging problem, our team comes up with the idea of this case. In our project, we intend to construct two plasmids to carry some specific elements and flocculation gene into the E.coli host，so that we can detect whether the environment is contaminated with heavy metal ions and report the concentration of them. Owing to the flocculation gene, we can collect the reconstructed bacteria just in case

of repeated pollution. We focus on trace contaminates detection and set the goal of standardizing the whole detection protocol.

- **Title：** Dioxin Detective Y

Yifan Wu *Harbin Institute of Technology*

Abstract： Dioxins, by-products of various industrial processes, are commonly regarded as highly toxic compounds that are environmental pollutants and persistent organic pollutants. Its deleterious hazards intensify with the odorless, colorless and fat-soluble properties, which enable dioxins to accumulate in vivo thus threaten almost the whole biosphere. Considering their ability of causing reproductive and developmental problems, damaging the immune system, interfering with hormones and also leading to cancer, hand in glove with the difficulty of being tested, dioxins are worthy of the name “poison of the century”.

So as to detect dioxins blisteringly, we have constructed device for rapid detection of dioxins inside yeast. We utilized associability between the gene for testing dioxins of MOUSE as receptor and dioxins as antibodies, together with the regulation conducted by lexA operator to downstream promoters so that mouse dioxin receptor and lexA are combined, and, striding a step further, induce downstream gene with the presence of dioxins. Eventually, we can achieve our goal of detecting dioxins rapidly through the system of positive feedback of fluorescence. Concerning the properties of yeast, like non-toxic and able to survive in hypertonic environment, our project can be put into wide application in the field like food detection, hence guarantee the food safety among the community of we human beings.